Expert Systems With Applications 62 (2016) 212–224

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Flash mobs, Arab Spring and protest movements: Can we analyse group identities in online conversations?

Natalia Criado[a,*], Awais Rashid[b], Larissa Leite[b]

[a] King's College London, London, WC2R 2LS, United Kingdom
[b] Security Lancaster, Lancaster University, Lancaster LA1 4YW, UK

ARTICLE INFO

Article history:
Received 10 February 2016
Revised 11 April 2016
Accepted 12 June 2016
Available online 16 June 2016

Keywords:
Social identities
Online social media
Natural language processing

ABSTRACT

The Internet has provided people with new ways of expressing not only their individuality but also their collectivity i.e., their group affiliations. These group identities are the shared sense of belonging to a group. Online contact with others who share the same group identity can lead to cooperation and, even, coordination of social action initiatives both online and offline. Such social actions may be for the purposes of positive change, e.g., the Arab Spring in 2010, or disruptive, e.g., the England Riots in 2011. Stylometry and authorship attribution research has shown that it is possible to distinguish individuals based on their online language. In contrast, this work proposes and evaluates a model to analyse group identities online based on textual conversations amongst groups. We argue that textual features make it possible to automatically distinguish between different group identities and detect whether group identities are salient (i.e., most prominent) in the context of a particular conversation. We show that the salience of group identities can be detected with 95% accuracy and group identities can be distinguished from others with 84% accuracy. We also identify the most relevant features that may enable mal-actors to manipulate the actions of online groups. This has major implications for tools and techniques to drive positive social actions online or safeguard society from disruptive initiatives. At the same time, it poses privacy challenges given the potential ability to persuade or dissuade large groups online to move from rhetoric to action.

1. Introduction

Global and national events over recent years have shown that online social media can be a force for good (e.g., Arab Spring in 2010) and harm (e.g., the England Riots in 2011). In both of these examples, social media played a key role in group formation and organisation, and in the coordination of the group's subsequent collective actions (i.e., the move from rhetoric to action) (Halliday, 2011; Tufekci & Wilson, 2012). Such coordinated actions are possible because individuals identify themselves with a particular social group or with an ideal (Taylor, Whittier, & Morris, 1992). Online identity in such contexts is, therefore, not so much about the categorisation of the self as a singular "I". Instead it is the conception and expression of group affiliations as a more inclusive "we".

This paper focuses on these online group identities. Offline group identities are usually referred to as *social identities* by social identity theory (Deaux, 1996; Stryker & Burke, 2000; Tajfel, 2010), a social psychological theory that sets out to explain group processes, intergroup relationships and the social self. Social identity is the individual's explicit or implicit expression of belonging to certain social group, together with some emotional and value significance to him/her of the group membership (Tajfel, 2010). Thus, a person has not one "personal self" but rather multiple social identities that are culturally contingent and contextual (Hankin, 2013). The *salient identity* is the identity that comes into play and is invoked in a specific situation or context (Stryker & Burke, 2000). Thus, a social identity is salient when it is invoked across a group of persons who perceive themselves as members of a social group. Which identity becomes salient in a given situation depends on factors such as the level of commitment of a person to a particular identity. One component of commitment is the number of others with whom one is connected by possessing a particular identity. Thus, when a person shares a certain identity with a greater number of people, his/her commitment to that identity tends to be higher and this identity is likely to be more salient (Stryker, 1980).

Given the importance of online social media in orchestrating and coordinating large-scale group mobilisations —from democracy

* Corresponding author.
E-mail addresses: natalia.criado@kcl.ac.uk (N. Criado), a.rashid@lancaster.ac.uk (A. Rashid), larissa.leite@gmail.com (L. Leite).

http://dx.doi.org/10.1016/j.eswa.2016.06.023
0957-4174/

and protest movements to hacktivist groups through to riots and extreme right wing marches— group identities are of key interest to a variety of stakeholders. They can be: mobilised as a resource for positive social change; studied to understand and counteract organised online actions that may compromise the safety and security of citizens; and even potentially be harnessed to build resilience in individuals and groups to limit the harmful effects of government or extremist efforts to disrupt online group formation and subsequent mobilisation.

Of course, group identities are not the only variable that predicts behaviour, but they can provide a guide to likely behaviours —as stated by social identity theory the higher the salience of a social identity (i.e., the identification with a particular group), the greater the individual's willingness to contribute to the social action (Bagozzi & Dholakia, 2002; Stryker & Burke, 2000). Social identities have been shown to influence behaviour in many domains, including politics (Jackson & Smith, 1999), protest movements (Reicher, 1996) and fan behaviour (Platow et al., 1999). Knowing how salient is group identity can lead to predictions of how much the identity will influence the individuals' beliefs, emotions and actions. Since the activation of a social identity affects the way people think as well as their feelings and behaviours, our hypothesis is that such group identities also affect the way in which people communicate online. As such, our model characterises text-based online communications in terms of a set of textual features such as their language, their style and their interaction patterns (i.e., the way in which users interact). We then study the features that can best distinguish between different group identities online as well as those features that can indicate the salience, or lack thereof, group identities. We address research questions categorised as follows:

1. Detecting salience of group identities:
   (a) Do group identities manifest in online conversations, i.e., is it possible to use textual features to automatically detect the presence of salient group identities?
   (b) Is our analysis model generalizable to detect identity salience across different group identities and on different online social media?
   (c) Which features are most suitable for detecting identity salience?
2. Distinguishing group identities:
   (a) Is it possible to distinguish between different group identities on the basis of textual features automatically extracted from conversations?
   (b) Is our analysis model generalizable to distinguish group identities over time and on different online social media?
   (c) Which features enable a specific group identity to be accurately predicted?

Our evaluation shows that, by using a range of structural, grammatical, semantic, categorical and stylistic features, our model can detect the salience of group identities with 95% accuracy and distinguish between group identities with 84% accuracy. In general, our study reveals that there is much more valuable information available on social media than just personal data. We identify features of online conversations that can reveal important dynamics of online groups and, hence, potential drivers for mobilisation of such groups. Notwithstanding the importance of protecting personal data on online social media (Anthonysamy, Greenwood, & Rashid, 2013; Madejski, Johnson, & Bellovin, 2011), it is also important to study and understand how group identities are formed and could be exploited for positive or negative ends. While the former has the potential to adversely affect individuals, the latter has major implications for social action/inaction in our modern digital society.

The novel contributions of this paper are fourfold:

1. This is the first paper to propose a model to analyse online group identities based on social identity principles.
2. We use textual features to detect group identity and its salience. In contrast with other works that study the difficulties people encounter when interacting with heterogeneous groups in an online social network, e.g., (DiMicco & Millen, 2007), or how online identities are constructed and shaped, e.g. (Zhao, Grasmuck, & Martin, 2008), all the features analysed in our model are extracted fully automatically, i.e., no human intervention is required.
3. We demonstrate that group identities and their salience manifest themselves, with a high degree of accuracy, in text-based online communications through a range of structural, grammatical, semantic, categorical and stylistic features.
4. Our results open up key privacy challenges for the research community at large with regards to the potential exploitation of group identities to persuade or dissuade large groups online to move from rhetoric to action. We have implemented an online tool that enables study of features underpinning online group identities in order to investigate these challenges. We identify which features can put online groups at most risk of such manipulation by mal-actors so as to build resilience against such out-group influences.

The rest of the paper is organised as follows: Section 2 describes related work. Section 3 presents our model for analysing group identities including the features and the classifiers used in the analysis. Section 4 describes experiments that validate of our model including the datasets used and the results obtained. We discuss the implications of our model and experiments in Section 5. Finally, Section 6 concludes the paper and identifies directions for future work.

2. Related work

Within the Artificial Intelligence field different computational models have been proposed to represent social identities. One of the most cited models is the ABIR (Agent-Based Identity Repertoire) model (Lustick, 2000), which seeks to refine, elaborate, and test theories of identity and identity shifts. This model has been used in agent-based simulations to analyse the emergence (Rousseau & Van Der Veen, 2005) and dynamics (Smaldino, Pickett, Sherman, & Schank, 2012) of social identities offline. To the best of our knowledge, ours is the first model for the automatic analysis of group identities invoked on different online social media.

There are empirical proposals, as ours, that draw conclusions about identity from information extracted from online social media. DiMicco and Millen (DiMicco & Millen, 2007) describe a study about the way in which people present themselves (i.e., the way in which people invoke their identities) on Facebook. Specifically, the authors analysed Facebook profiles and interviewed employees belonging to the same company with the aim of understanding how they managed their identity when interacting with different social groups (e.g., family, friends from school, workmates, etc.) on Facebook. The main contribution of their study was the identification of the difficulties that people encounter when interacting with heterogeneous groups using the same online social network; and the identification of the need for more sophisticated controls that help one to manage one's identities online. Similarly, Zhao et al. (Zhao et al., 2008) analysed Facebook profiles of students in a university to study how these students presented themselves on Facebook. They focused on how the online identities of these persons were "built" on Facebook. An interesting conclusion of their study is that identities are usually claimed implicitly on Facebook (e.g., people express that they belong to a group of friends by posting pictures with these friends instead of writing it in their self-description).

Our model, on the other hand, explores group identities by focusing on automatically analysing the interactions among users in which these identities are implicitly salient.

In recent study, Conover et al. (Conover et al., 2011) utilised clustering and manual annotation of tweets to analyse the way in which people with different political orientations (i.e., political identities) communicate on Twitter. Specifically, they analysed the retweets and mentions (which include replies) among users with different political orientations. Their study shows that tweets are usually retweeted by users who have a homogeneous political orientation. In contrast, tweets are mentioned by users with a heterogeneous political orientation.

In all the proposals aforementioned, the information is manually analysed and processed by humans. However, there are other proposals, like ours, in which the information is automatically analysed and processed. Research in the field of stylometry and authorship attribution has focused on automatically distinguishing between individuals online (Narayanan et al., 2012; Stamatatos, 2009) as well as deception detection in online conversations (Afroz, Brennan, & Greenstadt, 2012; Rashid et al., 2013). In contrast, our approach focuses on analysis of group identities instead of personal characteristics.

Within the area of Social Networks a significant amount of work has been done to detect user communities or densely connected subgroups of users in the network (Girvan & Newman, 2002). In particular, several tools have been proposed to detect communities automatically using unsupervised machine learning algorithms (Culotta, Bekkerman, & McCallum, 2004; Fogués, Such, Espinosa, & Garcia-Fornes, 2014; Matsuo et al., 2007). Although communities and group identities are not exactly the same concept (e.g., the fact that users belong to the same domain in a network does not entail that they feel as members of the same social group), it might be argued that the same techniques used for community detection can be used for group identity analysis. However, proposals on community detection assume that the social network graph is known (i.e., the users and the relationships between users are known), which makes possible the identification of tightly knit groups of users. However, this assumption is too strong for group identity analysis because it is not necessarily true that all users sharing a group identity are known. Similarly, user relationships are likely to be unknown in this prediction problem. For example, in many online social media, like Facebook, the information about users' friends is private and cannot be exploited to detect group identities that are expressed explicitly (i.e., by means of friendship relationships). This paper goes beyond these approaches by using textual features to analysis group identities that are implicitly salient in online conversations.

Recent research on Natural Language Processing is directing its efforts towards analysing short text messages exchanged online (Han & Baldwin, 2011). In particular, several authors have proposed to combine machine learning and natural language processing techniques to produce models that annotate short text messages with tags identifying specific themes or content (Ramage, Hall, Nallapati, & Manning, 2009). Note that these techniques can be used to detect conversations corresponding to specific topics, which could be used to perform group identity analysis. However, the fact that a conversation is associated with a cohesive set of topics does not necessary imply that the users share a common social identity. For example, messages posted by users in a given review site (e.g., TripAdvisor[1]) may be associated with a reduced set of topics according to the nature of the site (e.g., food, accommodation, attractions, etc.), but it is not necessarily true that these users identify themselves as members of the same social group. Besides that, these models only allow messages to be annotated with a predefined set of tags and, as a consequence, they will fail to detect unforeseen topics that may be associated with emerging group identities. Our research also combines machine learning with different analysis techniques such as NLP, stylometry and interaction analysis to produce a model specifically aimed at predicting group identities.

Verma et al. (Verma et al., 2011) propose and evaluate a classifier to detect those tweets that contribute to "situational awareness" in mass emergency. In (Gupta & Kumaraguru, 2012) the authors have built a linear regression model that takes as input text-content based features to predict the credibility of tweets. Similarly, in (Ratkiewicz et al., 2011) a web service is presented that automatically detects astroturfing (i.e., campaigns coming from disinterested, grassroots participants that are in reality carried out by a single person or organisation) in Twitter. In a more recent work, Cheng et al. (Cheng, Romero, Meeder, & Kleinberg, 2011) analyse the structural properties of social networks to predict reciprocity of communication among Twitter users. Similar to our approach, these proposals illustrate the potential information that can be gleaned by automatic analysis of online social media interactions.

In a recent proposal, Charitonidis et al. (Charitonidis, Rashid, & Taylor, 2015) analysed online communications to study offline group action processes. In particular, this work analysed different Twitter conversations for a specific event to identify weak signals that could be used to predict offline group actions. These results evidence that there are such early indicators of group actions in online communication. Based on these findings, our paper proposes a novel model to predict the salience of group identities in online conversations. In particular, our work complements and extends this research by allowing the automated detection and identification of group identities belonging to different domains and online media; as opposed to the work of Charitonidis et al. (Charitonidis et al., 2015) which analyses Twitter conversations corresponding to a specific event and does not propose a general predictive model.

3. Group identity model

In this section we present a formal description of the model used for analysing group identities. The aims of our model are twofold. Firstly, we aim to determine if we can automatically *detect* the existence of salient group identities in text-based online communications. Such automatic detection of identity salience would allow the detection of incipient and unforeseen group identities that might lead to social action —as mentioned earlier, several works on social identity theory have noted the potential causal relationship between the salience of a social identity and social action (Bagozzi & Dholakia, 2002; Stryker & Burke, 2000). Secondly, we aim to determine if we can automatically *distinguish* between different group identities in text-based online communications. This would enable automatic classification of interactions according to group identities of interest; e.g., group identities that may be considered as dangerous or beneficial.

3.1. Model overview

In our model each user u corresponds to an individual. We denote by $\mathcal{U}$ the set of users that communicate online. We also assume that there is a distinguished set $\mathcal{I}$ of group identities that correspond to social groups (e.g., supporter of Manchester United). Each user may belong to different social groups (i.e., s/he may have different group identities).

The information exchanged among users is formalised as tuples $\langle s, R, c\rangle$; where $s \in \mathcal{U}$ is the sender, $R \subseteq \mathcal{U}$ is the set of receivers,

[1] http://www.tripadvisor.com

and c is the message content. In this paper we only consider text-based messages. Thus, the content of a message c consists of an ordered set of words $\{w_1, \ldots, w_n\}$, and a set of terms $\{d_1, ..., d_k\}$ containing metadata.

In each message, the sender user can invoke one or more group identities. We define a function *invoked* that maps each message with the group identities that are invoked in it; i.e., given a message $\langle s, R, c\rangle$, the identities invoked in it are defined as $invoked(\langle s, R, c\rangle) \subseteq \mathcal{I}$.

Group identities are culturally contingent and contextual and, therefore, invoked in specific contexts or situations (Hankin, 2013). In online textual communication, a message sent by one user is usually replied by other users and the context or situation in which group identities are invoked is formed by related messages. In particular, we define a set of related messages (i.e., a message and its replies) as a conversation. More formally, we define a conversation as a set of ordered messages $\{m_1, ..., m_n\}$ where each message m_i is defined as a tuple $\langle s_i, R_i, c_i\rangle$. In a conversation, the first message (m_1) is a conversation initiation message, i.e., the message that started the conversation; and the rest of the messages ($\{m_2, ..., m_n\}$) are replies to this message.

As aforementioned, one of the key factors that make it more likely that a group identity is salient is the connectedness among persons who possess this particular group identity. Thus, when a person interacts with others by invoking a group identity and the others confirm this identity, then the salience of this identity is reinforced (Stryker & Burke, 2000). In fact, we hypothesise that this is one of the main reasons for the formation of online communities, to create situations in which group identities can be expressed and confirmed. The salient group identities in a conversation are the group identities that are invoked repeatedly across the messages in a conversation. More formally, we define the group identities that are salient in a conversation as:

$$salient(\{m_1, \ldots, m_n\}) = \bigcap^{\{q\}} invoked(m_i)$$

where $q \in \mathbb{N}$ such that $q < n$ and $\cap^{\{q\}}$ is the relaxed intersection of sets, which corresponds to the classical intersection between sets except that it is allowed to relax q sets in order to avoid an empty intersection. Thus, the salient group identities is the set of all group identities that are invoked across all the messages ($invoked(m_i)$), except q messages at most. Note that the relaxed intersection makes it possible to make a robust identification of salient identities in a conversation with respect to some outlier messages that invoke identities that are not predominant in the conversation.

When a conversation involves users who share the same group identity and are aware of it, it is highly probable that this group identity is invoked across most of the messages.[2] In contrast, when a conversation involves users who do not share a group identity or who are not aware of this fact, then it is highly probable that the messages in the conversation invoke disparate identities and, as a consequence, the salient group identity is the empty set. Accordingly, we define *salience* as a function that determines if a group identity is salient in a conversation as follows:

$$salience(\{m_1, \ldots, m_n\}) = \begin{cases} True & \text{if } salient(\{m_1, \ldots, m_n\}) \neq \emptyset \\ False & \text{otherwise} \end{cases}$$

[2] Note that it is also probable that a small proportion of the messages in a conversation are sent by users who belong to an opposite group and want to confront the users sharing the salient group identity, which, in turn, reinforces the salience of this group identity (Reicher, 1996).

3.2. Features analysed

We use linguistic and structural features of conversations to predict the values of the *salient* and *salience* functions. In particular, we analyse five feature sets that can be extracted from online conversations. These feature sets are further classified into three main categories: (i) online interaction patterns, (ii) natural language features, and (iii) stylistic metrics.

Online interaction patterns. This category includes features that can be extracted by analysing structural metrics of an online conversation:

Structural feature set. This set is formed by 3 numeric features: (i) the number of messages contained in a conversation; (ii) the participation-level of users, i.e., the ratio of users to the number of messages; and (iii) the average influence of messages; defined as the average number of likes or retweeted count of messages.

Natural language features. This category includes features that can be extracted by applying natural language processing techniques to the text of the messages contained in the conversations. Specifically, we make use of the techniques proposed by Rayson in (Rayson, 2008) to extract natural language features, since these techniques have been successfully used to analyse online conversations extracted from Peer-2-Peer networks (Hughes et al., 2008) or Twitter (Ferrario et al., 2012). These features are grouped into three feature sets:

POS feature set. This set is formed by numeric features that represent the relative frequency of basic parts-of-speech (POS) in the messages contained in conversations. Examples of such features include relative frequency of articles, adjectives, nouns, etc. To carry out the POS tagging, we use the CLAWS (Garside, 1987) tagger, which considers a tagset with 138 POS tags.

Semantic feature set. This set is formed by numeric features that represent the relative frequency of semantic tags in the messages contained in conversations. Examples of such semantic features include the relative frequency of text classified as "geographical names" or text pertaining to "groups and affiliations", etc. To carry out the semantic tagging, we use the USAS (Wilson & Rayson, 1993) system that considers a tagset with 452 semantic tags.

Category feature set. This set is formed by numeric features that represent the relative frequency of 36 categories or key concepts that may manifest in a conversation. These key categories are obtained by applying the keywords methodology (i.e., applying the keyness calculation to word frequency lists) to extract key domain concepts (i.e., applying the keyness calculation to semantic tag frequency lists). Examples of such features include relative frequency of categories such as sports, politics, etc. To identify categories in conversations we make use of Rayson's approach (Rayson, 2008).

Stylistic metrics. This category includes stylistic features that can be extracted by tools and methods from the field of authorship attribution (Stamatatos, 2009). Specifically, we use the stylistic metrics proposed in (Rashid et al., 2013), which have been used detecting masquerading behaviour online:

Style feature set. This set is formed by 22 numeric features. Examples include: the average length of messages in terms of words and characters, the frequency of emoticons, and the vocabulary richness.

Note that our feature sets only include features that can be analysed considering the information that is publicly available online. Other features such as demographic information about the

users interacting in conversations may be private and cannot be exploited to predict group identities.

3.3. Classifiers used

The features above are used to predict group identities in conversations. Specifically, the analysis is aimed to detect salient group identities and to identify group identities. To this aim we have built two types of classifiers:

- Detection classifiers, which classify conversations into two categories: *identity salience*, and *no salience*. A conversation (c) belongs to the category *identity salience* when there is a group identity that is salient in the conversation (i.e., when *salience*(c) is *True*); and to the category *no salience* otherwise.
- Identification classifiers, which classify conversations into a finite set of categories corresponding to the group identities that are salient in the conversations. More formally, given a conversation (c) an identification classifier tries to predict the value of the salient function for this conversation (*salient*(c)).

We have implemented each type of classifier using two different algorithms: a J48 classifier and a Support Vector Machine (SVM) classifier, using the implementations in the Weka[3] data-mining tool. These two classifiers have demonstrated a good performance on classification tasks with textual data (Afroz et al., 2012; Burgoon, Blair, Qin, & Nunamaker Jr, 2003). To train these classifiers we annotate each conversation in our dataset with its class and the values of the different features analysed. This leads to five training sets, one per feature set —in each training set conversations are annotated with their class and the values of one feature set.

3.4. Tool

We have implemented our model in a tool, *Identi-scope*, that makes these classifiers and the underlying feature extraction tools available as a workflow to support studies of group identities and the potential that may come from harnessing a deeper understanding of the processes that underpin such identities. The tool enables users to study features underpinning identities of groups that make their conversations available publicly. At the same time, users can point the tool to their private conversations to understand the various group identities they inhabit online and the features that underpin those identities. The analysis can be conducted over different time periods in the same conversation to study fluctuations of social identities in response to particular stimuli, for instance, when key features underpinning social identities are changed. However, we note that, due to ethical reasons, we have not introduced such stimuli into any conversations in order to study such fluctuations. They can be a useful tool for users or groups to study how their activities online may be influenced by actors aiming to persuade or dissuade them from specific actions.

4. Evaluation

4.1. Datasets used in evaluation

To evaluate our analysis model we collected text-based datasets from Facebook and Twitter. According to our model, we refer to each post, comment or tweet as a *message*. Thus, the content of the message is formed by textual content and the metadata contains information about the message influence (i.e., the number of likes in case of posts and comments, and the retweet count in case of tweets). Finally, the term *conversation* refers to a collection that includes: a text-based message (i.e., a tweet or post) and other related text-based messages (i.e., replies or comments).

4.1.1. Facebook datasets

All the information collected was publicly available on Facebook in two different periods: (i) between 18th February 2013 and 20th April 2013; and (ii) between 12th May 2014 and 12th June 2014. For simplicity, we will refer to these collection periods as first and second, respectively.

Recent work on social psychology (Levine & Koschate, 2014) has demonstrated that the Internet provides users with online spaces for expressing their social identities. Their study on *Mumsnet*,[4] a website for parents that hosts discussion forums focused on different topics (e.g., mums, feminism), demonstrates that different forums represent different social identities (i.e., feminist forums represent the feminist social identity) and that the individuals change their writing style to adapt towards the group norm when a social identity is salient (e.g., when they post on the feminist forums). In accordance with these results, we collected information from the Facebook pages of protest groups, sports teams and personalities to obtain information about conversations in which group identities are salient.[5] Specifically, we collected conversations posted during our two collection periods from the Facebook pages of Anonymous, Barack Obama, Beyoncé, Lady Gaga and Manchester United. These pages are means to achieve or maintain positive a public image and to build a social group around a given protest group, sports team or personality. Supporters in turn use these pages to express their affection and support and to communicate with other fans. Most of these messages invoke the group identity of being a supporter of a particular person, sports team, or protest group. Besides that, users who belong to opposition groups can occasionally post messages in these pages to confront the salient group identity. As highlighted by social identity theory (Reicher, 1996), these opposition messages reinforce the salient group identity even further.[6] Thus, we assume that the impact of outlier messages (i.e., messages invoking disparate group identities) in these conversations is negligible and that all conversations belong to the *identity salient* class. For example, we have collected posts and comments from the Facebook page of Manchester United Football Club. This page is used by someone on his behalf of Manchester United to post information about its activities. Besides, this page is used by thousands of users (mainly Manchester United supporters) who comment on the posts. These users hold a common identity (being a Manchester United supporter) and view themselves as members of the same social group. For example, one of the messages in our dataset posted by someone on behalf of Manchester United contains the following text “Rafael wins your Man of the Match vote for his fantastic display at both ends vs. QPR. Well done Rafael!”. Among the messages sent in response to this post we can find messages like “Oh! He deserves it”, “What goal Rafael !!!”, and “Yeah,congrats to rafael. And hope best for you”.

To obtain information about conversations in which group identities are unlikely to be salient, we focused on those situations in which a person interacts with others on sites where heterogeneous information is published neutrally; i.e., in pages that do not try to create a social group around a protest movement, personality or sports team. In particular, we obtained a dataset where identity salience is necessarily diluted by collecting a large number of

[3] www.cs.waikato.ac.nz/ml/weka/

[4] http://www.mumsnet.com

[5] Recall that social identities have been shown to relate to behaviour in these domains (Jackson & Smith, 1999; Platow et al., 1999; Reicher, 1996).

[6] Social identities become more salient in situations where a social group conflicts with a relevant opposition group (e.g., when the ideas or interests of opposite groups clash) (Reicher, 1996).

Table 1
Facebook pages included in our study.

Page	Collection Period	Influence	Importance
Anonymous	First	1025210	19635
	Second	1746674	24657
Barack Obama	First	35303430	1211796
	Second	41005573	879450
Beyoncé	First	43985350	424706
BBC World News	First	2950463	74426
Lady Gaga	First	55962915	210843
	Second	66155029	2029189
Manchester United	First	32299914	1356171
	Second	50249810	1628006
MTV	First	43297624	521441
NBC News	First	737559	76122
YouTube	First	71981268	464301

Influence=*Likes* Count
Importance=*Talking About* Count

Table 2
Facebook dataset (First collection period).

Page	C	M	U	W
Anonymous	6	1650	1410	45086
Barack Obama	142	94010	67419	4628148
Beyoncé	10	2946	2510	172923
BBC World News	612	52841	46044	1589676
Lady Gaga	4	1640	1304	21724
Manchester United	246	88434	77600	1547141
MTV	152	20765	19601	382797
NBC News	271	55616	44016	1314677
YouTube	33	1983	1938	24021

C=conversations, M=messages, U=Users, W=words

Table 3
Facebook dataset (Second collection period).

Page	C	M	U	W
Anonymous	4	452	410	11633
Barack Obama	7	3710	2902	143901
Lady Gaga	5	998	987	18260
Manchester United	39	8695	8446	190219

C=conversations, M=messages, U=Users, W=words

Table 4
Profiles included in our study.

Profile	Influence	Importance
Anonymous	911130	10334
Justin Bieber	36299641	548920
Barack Obama	28587127	189859
Lady Gaga	35229893	243017
Manchester United	538129	2265
MTV	7621274	25589

Influence=*Followers* Count
Importance=*Listed* Count

Table 5
Twitter API dataset.

Profile	C	M	U	W
Anonymous	201	824	714	12598
Barack Obama	219	2379	1529	41945
Justin Bieber	295	3289	2700	35870
Lady Gaga	5	64	47	894
Manchester United	730	1875	1814	30242
MTV	642	2122	1925	30481

C=conversations, M=messages, U=Users, W=words

conversations focused on varied topics and formed by messages with different tones. In such conversations, the social structures supporting the salience of group identities dissolve (Burke & Stets, 1999), which leads to users invoking disparate identities. For example, we have collected posts from news pages that aim to cover all social, political and other events fairly and impartially. Users reading and commenting on these news pages may have several group identities but the fact that they cannot warranty that a particular identity is shared by other users on the news pages makes the level of commitment to a particular group identity low and several group identities are likely to be invoked. Thus, we use these conversations as samples belonging to the *no salience* class. Specifically, we collected all the posts and comments made during the first collection period from the Facebook pages of BBC World News, MTV, NBC News and YouTube.

We collected information from these sources for three reasons. Firstly, they are highly influential and important (see Table 1). We define *influence* as the power to have an effect on other users. Accordingly, we define that a Facebook page is influential when many people like it. Thus, we consider the *likes* count as an influence measure. We define *importance* as the actual manifestation of effect on other users. Accordingly, we define that a Facebook page is important when many people mention it. Thus, we consider the *talking about* count as an importance measure. Secondly, these sources are frequently updated and commented on and, as a consequence, they contain a lot of information (see Tables 2 and 3). Finally, they cover different types of content, such as sports, politics, news, and so on.

4.1.2. Twitter datasets

Twitter API dataset. This dataset contains tweets publicly available on Twitter between 18 February 2013 and 20 April 2013 that have been collected using the Twitter public API.

Similar to our approach to the Facebook dataset, to obtain information about conversations in which a group identity is salient, we collected all the conversations from the Twitter profiles of Anonymous, Justin Bieber, Barack Obama, Lady Gaga and Manchester United. To obtain information about conversations in which group identities are unlikely to be salient, we collected tweets and replies from the Twitter profile of MTV.

We collected information from these sources because (see Tables 4 and 5): they are highly influential, important, contain lots of information and cover different types of content. We define that a Twitter profile is influential when many people follow it —i.e., *followers* count. Similarly, we define that a Twitter profile is important when many people list it —i.e., *listed* count (a list is a curated group of users).

2011 England Riots Dataset. This dataset contains the tweets exchanged during the 2011 England Riots. The riots are also called "BlackBerry riots" because people used mobile devices and social media to organise them (Halliday, 2011). Thus, this dataset contains real tweets exchanged during group identity formation processes, group identity invocation and social action coordination. Specifically, the disturbances reflected in our dataset began on Saturday 6 August 2011, after a protest in Tottenham following the death of Mark Duggan, a local who was shot dead by police on Thursday 4 August 2011. In the following days the riots spread across other parts of London and other cities in England including Birmingham, Bristol, and Manchester.

To collect this dataset, we have used Topsy,[7] which is a search engine for social posts and socially shared content, primarily on Twitter. The results provided by this search engine are not organised following a conversation pattern (i.e., an initiating tweet and

[7] http://topsy.com/

Table 6
Twitter England riots dataset.

Set	C	M	U	W
TottenhamPreRiots	29	3467	3151	60875
TottenhamRiots	29	4244	4203	63189
LondonRiots	29	4706	4699	73175

C=conversations, M=messages, U=Users, W=words

its replies). Thus, we approximated the conversations by grouping the tweets according to the time when they were exchanged (e.g., consecutive tweets belong to the same conversation). Specifically, we have selected a subset of the tweets corresponding to the riots in Tottenham and London as follows:

1. *TottenhamPreRiots*. This set contains tweets that match the query `#tottenham OR tottenham` and were exchanged on 4 Aug. 2011. At that point in time, the riots had not started in Tottenham and we assume that tweets invoke disparate identities and that conversations belong to the *no salience* class. In fact, during this period of time, the number of tweets matching this query per hour was lower than 500. Among these tweets we can find messages like: "has delighted the board of Tottenham Hotspur by winning the Premier Division" and "So we just got to tottenham hale & realised we left our money at home. Doh! Back we go".
2. *TottenhamRiots*. This set contains tweets that match the query `#tottenham OR tottenham` and were sent on the 6 Aug. 2011. At that point in time, riots were very prominent in Tottenham and, as mentioned above, Twitter and other social networks were used to coordinate social action. Indeed, during this period of time the number of tweets matching this query per hour was higher than 14000. Thus, we assume that group identities are salient in these conversations. Among these tweets we can find messages like: "It's not just #tottenham. #MetPolice = corrupt dishonest + unaccountable ie rotten to the core" and "I'm proud of tottenham right now".
3. *LondonRiots*. This set contains tweets that match the query `#londonriots OR (#london AND riots)` and were sent on 8 Aug. 2011. At this point, the riots were very prominent in London and the number of tweets matching this query per hour was $>$ 60,000 (compared to $<$ 2500 on the previous day). Thus, we assume that group identities are salient in these conversations. Among these tweets we can find messages like: "A riot is the language of the unheard - Martin Luther King. #londonriots" and "What army do we have to bring in - the majority of them are being used as target practice by the Taliban...!!! #londonriots".

Note the queries used for selecting the different datasets have been previously used to identify weak signals of real-world mobilisations in (Charitonidis et al., 2015).

Table 6 shows the number of conversations, messages, users and words contained in the riots sets.

4.2. Detecting salience of group identities

4.2.1. Do group identities manifest in online conversations, i.e., is it possible to automatically detect the presence of salient group identities online?

Our first research question is related to the detection of the existence of salient group identities in online conversations. To answer this question, we used the conversations extracted during the first collection period from the pages of Anonymous, Barack Obama, and Manchester United as examples of conversations where there are salient group identities. We used the conversations collected during the first collection period from the pages of BBC World News, NBC News and YouTube, as examples of conversations where there is no apparent salient identity shared by the users. We trained the J48 and SVM *detection classifiers* with the conversations annotated with each feature set. To assess the accuracy of these classifiers we used leave-one-out cross-validation —using a single conversation from the set as the validation data, and the remaining conversations as the training data; this was repeated such that each conversation in the dataset was used once as the validation data.

Table 7
Identity salience detection when different feature sets are used to train the classifiers.

Feature Set	Classifier	Accuracy(%)	ROC Area
Structural	J48	98.94	0.98
	SVM	89.71	0.86
POS	J48	96.98	0.96
	SVM	98.61	0.98
Semantic	J48	97.31	0.98
	SVM	97.96	0.97
Category	J48	96.73	0.97
	SVM	94.37	0.9
Style	J48	89.06	0.9
	SVM	81.71	0.68

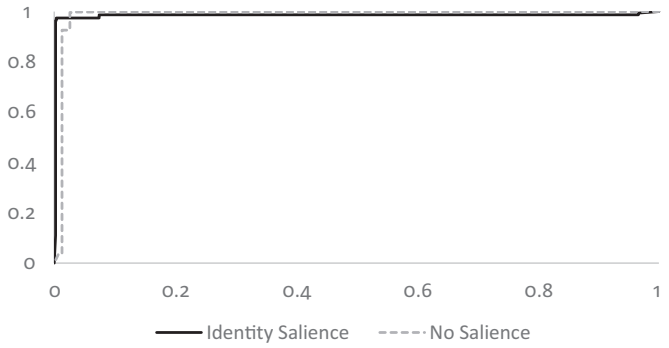

Fig. 1. ROC curves obtained for the identity salience detection problem with the J48 classifier and the structural feature set.

Table 7 shows the results obtained by each classifier when the structural features, POS features, semantic features, category features and style features are considered. Specifically, this table shows the accuracy, which is the percentage of correctly classified conversations; and the weighted (by class size) area under the ROC curve. Accuracy provides an understandable measure for classifier performance. However, accuracy must be interpreted with caution when classes in the dataset are unbalanced (as occurs in our experiments). In this situation, the area under the ROC curve is a more robust performance measure (Metz, 1978). According to guidelines for the interpretation of the area under the ROC curve, excellent classifiers obtain areas under the ROC curve within the interval (0.9, 1], good classifiers (0.8, 0.9], fair classifiers (0.7, 0.8], poor classifiers (0.6, 0.7], and fail classifiers obtain areas lower or equal to 0.6.

From the results in Table 7, we can determine that it is possible to detect the presence of salient group identities in online conversations with a high degree of accuracy —an accuracy of 98.94% is obtained with the J48 classifier and the structural feature set. Fig. 1 shows the ROC curves obtained by this classifier.[8] However, all the feature sets, with the exception of style features, allow the detection of identity salience with a high degree of accuracy, i.e., there

[8] To dismiss overfitting problems we also repeated this experiment using cross-validation, which is a well-known technique to avoid overfitting, with different numbers of folders and we obtained very similar results.

Table 8
Identity salience detection: Facebook generalization.

Feature Set	Classifier	Accuracy(%)	ROC Area
Structural	J48	75.3	0.74
	SVM	84.34	0.62
POS	J48	69.28	0.71
	SVM	66.27	0.82
Semantic	J48	83.13	0.91
	SVM	92.17	0.89
Category	J48	65.06	0.58
	SVM	46.39	0.61
Style	J48	89.76	0.93
	SVM	96.99	0.82
250 Features	J48	70.48	0.79
	SVM	89.16	0.94

is at least one classifier with an accuracy greater than 96% and the area under the ROC curve greater than 0.9.

We observe that the style feature set is less discriminative, e.g., the area under the ROC curve obtained by the best classifier trained with the style feature set is the lowest among all classifiers. This can be explained by the fact that the style features provide a characterisation of the style of the different persons but are not general enough to detect the common features that characterise the existence of salient group identities.

4.2.2. Are these results generalizable to detection of identity salience for unknown group identities and on different online social media?

Facebook generalization. In this experiment we aim to determine whether the previous results are generalizable to detect salience of unknown group identities. To this aim, we tested the classifiers with a different dataset extracted from Facebook. Specifically, we used the conversations collected during the first collection period from the Facebook pages of Beyoncé and Lady Gaga as examples of conversations where there are salient group identities. We used the conversations collected during the first collection period from the Facebook page of MTV as examples of conversations where there is no apparent salient group identity. Thus, we are evaluating if the classifiers are able to detect the salience of group identities that belong to different domains (i.e., the training set contains conversations about politics, sports, news, videos, whereas the test set contains conversations about music and TV).

Table 8 shows the results obtained. In general the accuracies of all classifiers are lower than in the previous experiment. Therefore, in order to determine which specific features are most suitable for detecting identity salience, we analysed the information gain (Kent, 1983) of features. The Information gain (IG) is frequently used in machine learning to define a preferred sequence of features to be used by a decision tree (such as the J48 classifier). Usually a feature with high IG should be preferred to other features. We calculated the IG of each feature for detecting identity salience and trained our J48 and SVM classifiers with the top 250 features by IG. This represents less than 40% of all 651 features arising from the union of all our feature sets. We used the same training data set as in Section 4.2.1 and tested the classifiers for detecting identity salience or otherwise using conversations from the pages of Beyoncé, Lady Gaga and MTV.

As can be seen in Table 8, the classifier trained with the most relevant features performs very well in this generalization in terms of accuracy and the area under the ROC curve. Specifically, it outperforms all classifiers trained with one feature set (i.e., an area under the ROC curve of 0.94 is obtained with the SVM classifier). Fig. 2 shows the ROC curves obtained by this classifier. This indicates that a combination of the high IG features from the various feature sets allows the most generalizable classifier to be trained. Therefore, we can conclude that, by using a combination of features, it is possible to automatically detect the salience of unknown group identities within the same online media (as used for training the classifiers) with a high degree of accuracy.

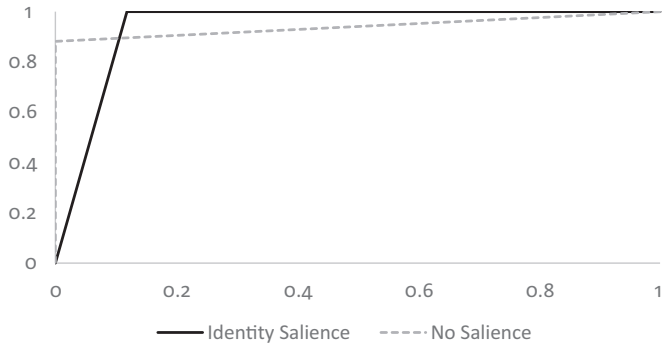

Fig. 2. ROC curves obtained for the identity salience detection problem when conversations from other Facebook pages are used to test the SVM classifier trained with the most relevant feature set.

Table 9
Identity salience detection: Twitter generalization.

Feature Set	Classifier	Accuracy (%)	ROC Area
Structural	J48	94.73	0.96
	SVM	73.89	0.73
POS	J48	80.15	0.82
	SVM	85.58	0.85
Semantic	J48	81.55	0.83
	SVM	85.17	0.85
Category	J48	75.29	0.77
	SVM	80.23	0.8
Style	J48	77.51	0.75
	SVM	72.41	0.69
250 features	J48	95.06	0.95
	SVM	90.69	0.9

Twitter Generalization. In this experiment we analyse if our analysis model is generalizable to detect identity salience on a different online media. We used the conversations contained in the *TottenhamRiots* set, in the Barack Obama and in the Justin Bieber profiles as examples of conversations where there are salient group identities; and conversations in the *TottenhamPreRiots* set and the MTV profile, as examples of conversations where there is not a salient group identity. Again, we employed leave-one-out cross-validation.

Table 9 shows the results obtained by the detection classifiers in this experiment. If we compare these results against the results of the Facebook detection experiment (described in Section 4.2.1), we observe that the performance of all classifiers deteriorates slightly in this experiment. This may be due to the fact that the tweet size is limited to 140 characters and fewer words are used to analyse conversations and train classifiers. As a consequence, the J48 classifier trained with the structural feature set (features not affected by the number of words in conversations) outperforms the rest of classifiers. Specifically, the J48 classifier trained with the structural feature set obtains an accuracy of 94.73% and an area under the ROC curve of 0.96 — see Fig. 3 for the ROC curves obtained this classifier. This demonstrates that our analysis model is generalizable to detect identity salience on different online media with a high accuracy. We can also observe from the table that both the J48 and SVM classifiers trained with the 250 most relevant features have a high accuracy —90% or above with a high ROC area ($\geq$ 0.9).

As in the Facebook detection experiment (described in Section 4.2.1), the style feature set is the less discriminative, i.e., the best classifier trained with the style feature is just a fair classifier since the area under the ROC curve is lower than 0.8. This

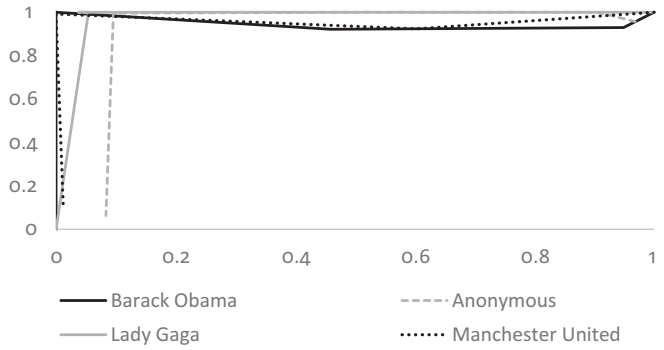

Fig. 3. ROC curves obtained for the identity salience detection problem with the J48 classifier and the structural feature set.

Table 10
Ten most relevant features for detecting identity salience.

Feature Type	Feature description	IG
Structural	Av. Influence	0.68
Structural	Messages	0.48
Semantic	General And Abstract Terms	0.37
Semantic	Measurement	0.36
Semantic	Social Actions, States And Processes	0.35
Semantic	Money generally	0.34
POS	Base form of lexical verb (e.g., give)	0.34
Semantic	Degree (i.e., intensifier terms)	0.34
Semantic	Quantities	0.33
Style	Av. Typing	0.32

supports our hypothesis that style features are not general enough to detect the common features of salient group identities.

4.2.3. Which features are most suitable for detecting identity salience?

Of the 250 most relevant features for detecting identity salience, 3 are structural, 110 POS, 104 semantic, 30 categories and 3 style. However, when we inspect these features in more detail we note that the IG of most individual features is not so substantially high to indicate that those features individually are strong indicators of identity salience. But our generalization experiments show that collectively they provide a strong basis for predicting identity salience.

Table 10 shows the 10 most relevant features for detecting identity salience. Specifically, for each relevant feature, it presents its type, description and IG value. We can observe that the IG values are low, which means that, in general, the features are less discriminative. Six of the ten most relevant features are semantic tags, two of them are structural features while one each is from the POS and style sets. This information is in line with the results achieved by the classifiers trained with the different sets of features (i.e., the three best classifiers in Section 4.2.1 are trained with the structural, POS and semantic feature sets). Specifically, the average IG of structural features is noticeably higher than the average IG of any other feature set. This explains the fact that the best classifier trained with the semantic feature set, which contains 6 of the 10 most relevant features, does not lead to better performance when compared with the best classifier trained with the structural feature set.

We can also observe that the two most relevant features are the structural features. This is explained by the fact that structural features allow to detect interaction patterns that characterise all group identities. For instance, it is possible that users who share a group identity are more prone to like comments that invoke this identity. However, detecting identity salience using structural features only may lead to poor results when detecting incipient group identities; such incipient group identities may have little influence. Furthermore, the structural feature of "average influence" has a high IG which reflects identity theorists' view of a cause-effect relationship, whereby the salience of a social identity influences collective action.

Table 11
Distinguishing group identities using different feature sets.

Feature Set	Classifier	Accuracy (%)	ROC Area
Structural	J48	95.21	0.98
	SVM	87.86	0.88
POS	J48	95.85	0.98
	SVM	98.4	0.98
Semantic	J48	98.72	0.99
	SVM	99.68	1.0
Category	J48	94.57	0.99
	SVM	98.08	0.99
Style	J48	91.69	0.97
	SVM	94.57	0.95

Fig. 4. ROC curves obtained for the identity identification problem with the SVM classifier trained with the semantic feature set.

4.3. Distinguishing between group identities

4.3.1. Is it possible to distinguish between different group identities?

Having determined that it is possible to detect the salience or lack thereof group identities in online conversations, we focus on the question of whether it is possible to distinguish between different group identities. Thus, in this case the class of each conversation is its group identity (i.e., Facebook page from which each conversation has been extracted). To answer our research question we used the conversations collected during the first collection period from the Facebook pages of Anonymous, Barack Obama, Lady Gaga and Manchester United to train *identification classifiers*.

From the results in Table 11, we can determine that it is possible to distinguish between group identities with a high degree of accuracy —an accuracy of 99.68% is obtained with the SVM classifier and the semantic feature set. Fig. 4 shows the ROC curves obtained by this classifier[9]. However, all feature sets allow group identities to be predicted with a high confidence (i.e., for all feature sets there is at least one classifier that obtains an area under the ROC curve greater than 0.9). We next analyse if these results are generalizable.

[9] Again, to dismiss overfitting problems we also repeated this experiment using cross-validation and we obtained very similar results.

Table 12
Distinguishing group identities: Time generalization.

Feature Set	Classifier	Accuracy (%)	ROC Area
Structural	J48	34.55	0.54
	SVM	78.18	0.62
POS	J48	72.73	0.75
	SVM	83.64	0.86
Semantic	J48	12.73	0.5
	SVM	60.0	0.7
Category	J48	45.45	0.52
	SVM	49.09	0.67
Style	J48	50.91	0.57
	SVM	69.09	0.6
250 Features	J48	80.0	0.74
	SVM	76.36	0.82

4.3.2. Are these results generalizable to distinguishing group identities over time and on different online social media?

Time generalization. We analyse whether the results in Section 4.3.1 can be generalized to distinguishing between group identities over time. It is obvious that classifiers can only classify instances into those classes that belong to the training set. This entails that we cannot use the classifiers to predict other group identities not included in the training set. Because of this, we can only determine if the results obtained in the above analysis, can be generalized to conversations invoking the same group identities over a different period of time. Note that the groups may evolve over time; e.g., the issues that are of interest to a group may change throughout time, but the collective sense of belonging to a specific group (i.e., the group identity) remains.

We test the accuracy of our previously trained classifiers, trained with the conversations collected during the first collection period from the Facebook pages of Anonymous, Barack Obama, Lady Gaga and Manchester United to distinguish between group identities in the conversations collected during the second collection period from the same pages. Thus, we are evaluating if the classifiers are able to predict the same group identities (i.e., being a supporter or opponent of Anonymous, Barack Obama, Lady Gaga or Manchester United) when they are invoked more than one year later. Similar to identity salience detection, we also determine the IG of all the features in the union of our feature sets and use the top 250 features to train J48 and SVM identification classifiers.

As we can observe from Table 12, the accuracy and the area under the ROC curve for all classifiers based on individual feature sets decrease. This is explained by the fact that these groups have evolved in terms of content (e.g., the main topics discussed in conversations), style (e.g., the number of words per message) and structure (e.g., number of users per conversation). For example, the influence and importance of Barack Obama's Facebook page increased noticeably between the two collection periods (see Table 1). As a result, it is possible that the number of users interacting in the page, the number of likes received per each message and the activity of these users changed drastically. These differences make it more challenging to predict group identities over time.

Despite these changes, we can observe that group identities can be identified with a high accuracy by the classifier trained with the POS feature set —an accuracy of 83.64% and an area under the ROC curve of 0.86 is obtained with the SVM classifier and the POS feature set. This entails that despite the passing of time, there are syntactic characteristics of each group identity that remain unaltered. We also observe that the classifiers trained with the 250 most relevant features perform better than those classifiers trained with individual feature sets (except the POS feature set). This shows that it is possible to use a combination of features to automatically predict and distinguish group identities over time with a high degree of accuracy —an accuracy of 82.92% is obtained with the SVM classifier and the most relevant feature set with substantial confidence in the prediction (i.e., the ROC area is 0.82). Fig. 5 shows the ROC curves obtained by the SVM classifier and the 250 most relevant features.

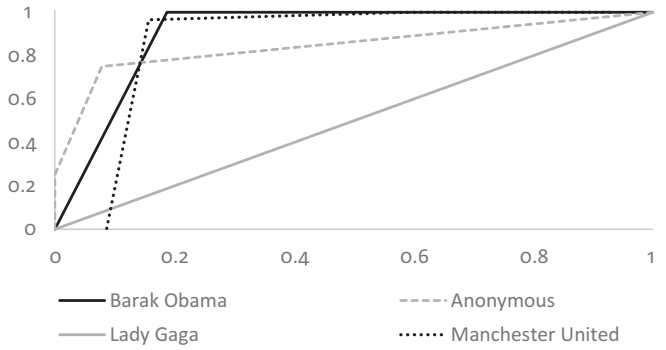

Fig. 5. ROC curves obtained for the identity prediction problem when conversations from a different time period are used to test the SVM classifier trained with the most relevant feature set.

Table 13
Distinguishing group identities: Twitter generalization.

Feature Set	Classifier	Accuracy (%)	ROC Area
Structural	J48	80.69	0.87
	SVM	71.37	0.74
POS	J48	75.74	0.79
	SVM	77.64	0.77
Semantic	J48	77.39	0.77
	SVM	82.84	0.86
Category	J48	72.85	0.72
	SVM	75.33	0.73
Style	J48	72.44	0.39
	SVM	70.71	0.65
250 features	J48	82.92	0.82
	SVM	87.62	0.89

Twitter generalization. We evaluate the extent to which our analysis can be generalized to identify group identities on different online media. To this end, we trained *identification classifiers* with the conversations contained in the *TottenhamRiots* and *LondonRiots* sets and the Twitter profiles of Barack Obama, Lady Gaga, Manchester United and Anonymous. Again, we employed leave-one-out cross-validation to assess each classifier and feature set.

Table 13 shows the results obtained. If we compare these results against the results of the Facebook identification experiment (described in Section 4.3.1), we observe that the performance of all classifiers deteriorates slightly in this experiment. Again, this may be explained by the tweet size limitation. Besides that, we have used conversations invoking similar group identities (i.e., group identities that correspond to different protest groups), which may be more difficult to distinguish from one another. Despite these similarities, group identities can be identified with high precision (i.e., there are classifiers that obtain areas under the ROC curve greater than 0.8). We observe that the classifiers trained with the 250 most relevant features perform better overall than those trained with individual feature sets. This shows that that our analysis model is generalizable to identify group identities on different online media with a high degree of accuracy – an accuracy of 87.62% is obtained with the SVM classifier and the most relevant feature set with substantial confidence in the predictions (i.e., the ROC area is 0.89). Fig. 6 shows the ROC curves obtained by the SVM classifier and the 250 most relevant features.

The classifiers trained with the style features obtain the lowest accuracies and ROC areas. This result confirms that style features

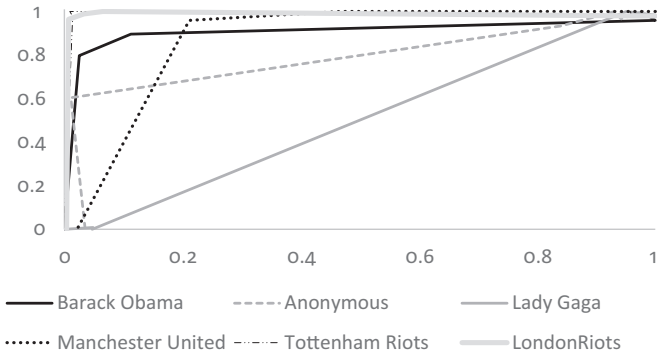

Fig. 6. ROC curves obtained for the identity prediction problem with the SVM classifier trained with the 250 most relevant features.

Table 14
Ten most relevant features for distinguishing group identities.

Feature Type	Feature description	IG
POS	Singular letter of the alphabet (e.g., a)	0.96
POS	Formula	0.92
Style	Av. Message Length (Chars)	0.92
Semantic	Power relationship	0.91
Semantic	General And Abstract Terms	0.9
Category	Sports	0.89
Style	Av. Message Length (Words)	0.88
Semantic	Measurement	0.87
Category	Crime	0.87
Semantic	Social Actions, States And Processes	0.86

provide a characterisation of the individual styles and are not general enough to distinguish the common features that characterise each group identity.

4.3.3. Which features enable a specific group identity to be accurately predicted?

Of the 250 features used to train the classifiers in Tables 12 and 13, 101 are semantic features, 106 POS, 35 categories, 5 style and 3 are structural features. This may lead one to conclude that semantic and POS features are the most necessary to characterise a group identity. However, when we study the IG of the top 10 features for predicting group identities (see Table 14), we observe that this is not necessarily the case. As illustrated by this table, 4 of the 10 most relevant features are semantic tags, 2 are style metrics, 2 are categories and 2 are POS tags; and all have very high IGs.

Interestingly, two style features show high IG, yet the best classifier trained with the style feature set is less accurate when compared with the best classifiers trained with the other feature sets. This is explained by the fact that the two high IG style features provide the same information (i.e., the more words in a message, the more characters it has) and their "addition" does not provide more information.

5. Discussion

Our results validate our hypothesis that group identities affect the way in which people communicate online and that it is possible to define a model that automatically analyses group identities using features extracted from text-based online communications. We now discuss some of the key insights and their potential implications.

Group identity manifests in semantic features. When distinguishing between group identities in Facebook (cf. Table 14), it is not surprising to note the presence of categories such as *Crime* and *Sports* —these are evocative topics and have been shown to have a causal connection with social identity formation (Levine & Crowther, 2008). The more interesting data is the presence of semantic features: *General and Abstract Terms*, which pertain to language use with regards to action/inaction in general, and *Social Actions, States and Processes*, which cover language use involving reciprocity, participation, friendliness and approachability. Interestingly these features also appear in the ten most relevant features for identity salience detection in Table 10 (albeit with a significantly lower IG). This reflects that formative processes for social identity manifest themselves in the semantics of the group conversations and can act as potential indicators for the emergence of social identities in online groups.

We also note the presence of the *Power Relationship* semantic feature in Table 14. This feature covers terms depicting power/authority/influence and organisation/administration. Also noteworthy are: the structural indicator of *average influence* and semantic indicator of *intensifier terms* (depicted by the *Degree* semantic tag) for identity salience in Table 10. Together, these point to a potential link between such features and social identity and group mobilisation. All these features merit further investigation.

Impact of the type and nature of social media. Our attempts at generalizing our analysis show reasonably high degrees of accuracy. However, they also indicate that the very nature of the social network and that of the data it carries has an effect. In particular, our generalization experiments show that there is not a single feature set that is able to produce satisfactory results in Facebook and Twitter. This may be attributed to the limits on message size in Twitter or how identity is implicitly expressed on different social media owing to the different features they afford to users (Conover et al., 2011; Zhao et al., 2008) and how these features may lead to various "in-group" and "out-group" formulations. However, the 250 most relevant features have been able to predict group identities with high accuracies across different social media. It would be interesting to study how other types of social media impact the accuracy of such a predictive approach and whether a hybrid feature set drawing upon training data from a range of online social media can provide a basis for accurately detecting incipient group identities.

The ethics of it all. The possibility of automatically predicting group identities poses a broad range of challenging ethical questions. For example, the features analysed in our study may be used for monitoring the evolution of group identities over time. This may permit the identification of different steps involved in the consolidation of social identities online. In turn, it may be possible to identify actions (e.g., shifts in behaviour) that reliably impact on a group's subsequent behaviour. By "seeding" specific semantic or structural features in text-based communications it may be possible to make specific identities salient and hence "nudge" the group's behaviour towards a specific outcome. On the one hand, democratic movements such as the Arab Spring could be promoted; e.g., by creating messages with an strong emphasis on reciprocity, participation, friendliness and approachability.[10] On the other hand, however, so could be violent actions such as the England Riots. These questions are highly pertinent given recent high profile news of mass surveillance activities such as Prism and the Snowden leaks.

We have implemented the Identi-scope tool that can enable exploration of these challenges. Furthermore, groups can utilise the tool to study if their conversations are being systematically nudged towards particular action or inaction through manipulation of the features we have identified.

[10] According to our experiments, these topics are key features underpinning group identities.

Limitations of our analysis. Our evaluation and results are based on data collected from Facebook and Twitter. As we note above, the nature of the social network and the purpose for which it is used by the various parties involved can influence the way social identities manifest themselves. Data from other social networks, especially those that cater for specific demographics, e.g., young people, or particular group affiliations (political, religious, etc.) may yield different results. One may conjecture that social networks that are aimed at particular group affiliations are likely to yield more accurate prediction of group identities. At the same time, there may be more fine-grained social identities at play (compared to the coarse-grained group identities in our study) in these social networks. Further experimentation is needed to determine whether the automated analysis presented in this paper will yield high accuracies for such fine-grained social identities.

Our model predicts those group identities that are sustained by online interactions. Notwithstanding the role of offline interactions in social identity formation and processes, our model only considers the information that is publicly available in online social media to predict group identities. The creation of a hybrid model capable of considering both online and offline interactions when predicting social identities is left as future work.

Model feasibility. The proposed analytical model and the *Identiscope tool* make extensive usage of different APIs provided by third parties. In particular, both the Facebook Graph API[11] and the Twitter public API[12] are used to collect the conversations to be used by the analytical model. Note these two APIs impose rate call limits to non-paying users. Similarly, our model and tool also make use of the Relative Insight's Interaction Analysis Engine[13] for NLP tasks, which is also accessed via its API over the Internet. Thus, the time need to analyse messages may be affected by network latency, congestion etc. However, these APIs process a reasonable amount of requests in a short period of time (e.g., the Facebook API allows us to collect information about 1000 messages in less than 6 seconds). Finally, we would like to mention that, for those domains where the volume and speed at which data is produced makes it necessary to reduce the time needed for processing, solutions such as parallelisation of the analysis, usage of paid APIs, etc. can be applied.

Model vulnerabilities. Our experiments demonstrate that our model is robust to predict group identities even if there are messages that invoke outlier identities (i.e., we have not performed any preprocessing on conversations to filter out outlier messages). However, this robustness may not hold when evasion techniques are used to mask group identities. For example, Islamic State supporters could try to misdirect group identity detection by injecting into their conversations messages in which a fake identity is invoked. Even more, automated approaches could be envisioned so that a single entity (whether individual or organisation) controls a large number of fake accounts to launch such evasive attacks. However, these threats can be mitigated using existing sybil defences (Alvisi, Clement, Epasto, Lattanzi, & Panconesi, 2013; Fong, 2011), classification techniques (Thomas, McCoy, Grier, Kolcz, & Paxson, 2013), and stylometry techniques (Afroz, Islam, Stolerman, Greenstadt, & McCoy, 2014; Brennan, Afroz, & Greenstadt, 2012; Ding, Fung, & Debbabi, 2015) to discard fake accounts and messages. The specific mitigation techniques to be applied in a given situation may depend not only on the evasion techniques used by attackers but also on the nature of the social media[14].

[11] https://developers.facebook.com/docs/graph-api

[12] https://dev.twitter.com/rest/public

[13] https://relativeinsight.com/

[14] An study to generate specific mitigation strategies is beyond the scope of this paper.

6. Conclusion

The model and results presented in this paper provide a stepping stone towards understanding how group identities and their salience manifest in text-based communications via online social media and the implications this holds regarding risk posed by external agents (government or otherwise) to influence collective action/inaction mediated by online social media. Our results show that it is possible to use linguistic and structural features and machine learning techniques to automatically distinguish between specific group identities as well as detect when group identities may be salient. Such predictions are not just highly accurate within a particular social networking platform but also show potential for generalization on different types of social networks. Particularly insightful are our observations about specific semantic features of the language used in conversations that indicate social processes for group formulation at work. Our analysis also shows a potential link between group identities and mobilisation inherent in the language of online groups. We also highlight the challenging ethical questions raised by the ability to detect and, potentially affect, social identities and their salience through analysis and manipulation of language features. We have developed a tool that allows exploration of social identities by individuals and groups so that they may develop resilience against outside agents attempting to influence their actions through manipulation of the features we have identified.

Our future work will focus on exploring specific research questions around the manifestation of particular types of semantic features and the impact of the nature of the social network as well as communication modes and processes on group identity analysis. Only by gaining a deeper understanding of the features and communication processes at play can we hope to unravel the various ethics and privacy questions raised by this paper.

Acknowledgements

This work has been carried out using Relative Insight's[15] Interaction Analysis Engine. The authors would like to thank Dr. Phil Greeenwood for his support with Interaction Analysis Engine, and Christos Charitonidis for giving us access to the riot dataset.

This work was partially supported by the EPSRC under grant "Identi-scope: Multiple identities as a resource for understanding and impacting behaviours in the digital world" (EP/J005053/1).

References

Afroz, S., Brennan, M., & Greenstadt, R. (2012). Detecting hoaxes, frauds, and deception in writing style online. In *Proceedings of the ieee symposium on security and privacy* (pp. 461–475).

Afroz, S., Islam, A. C., Stolerman, A., Greenstadt, R., & McCoy, D. (2014). Doppelgänger finder: Taking stylometry to the underground. In *Ieee symposium on security and privacy* (pp. 212–226).

Alvisi, L., Clement, A., Epasto, A., Lattanzi, S., & Panconesi, A. (2013). Sok: The evolution of sybil defense via social networks. In *Ieee symposium on security and privacy* (pp. 382–396).

Anthonysamy, P., Greenwood, P., & Rashid, A. (2013). Social networking privacy: Understanding the disconnect from policy to controls. *IEEE Computer, 46*(6), 60–67.

Bagozzi, R. P., & Dholakia, U. M. (2002). Intentional social action in virtual communities. *Journal of Interactive Marketing, 16*(2), 2–21.

Brennan, M., Afroz, S., & Greenstadt, R. (2012). Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security, 15*(3), 1–22.

Burgoon, J. K., Blair, J., Qin, T., & Nunamaker Jr, J. F. (2003). Detecting deception through linguistic analysis. In *Intelligence and security informatics* (pp. 91–101). Springer.

Burke, P. J., & Stets, J. E. (1999). Trust and commitment through self-verification. *Social Psychology Quarterly, 62*(4), 347–366.

Charitonidis, C., Rashid, A., & Taylor, P. J. (2015). Weak signals as predictors of real-world phenomena in social media. In *International conference on advances in social networks analysis and mining* (pp. 864–871). ACM/IEEE.

[15] https://relativeinsight.com/

Cheng, J., Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Predicting reciprocity in social networks. In *Proceedings of the ieee international conference on social computing* (pp. 49–56).
Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., & Menczer, F. (2011). Political polarization on twitter. In *Proceedings of the international conference on weblogs and social media*.
Culotta, A., Bekkerman, R., & McCallum, A. (2004). *Extracting social networks and contact information from email and the web* (pp. 1–8).
Deaux, K. (1996). *Social identification*. Guilford Press.
DiMicco, J. M., & Millen, D. R. (2007). Identity management: multiple presentations of self in facebook. In *Proceedings of the international acm conference on supporting group work* (pp. 383–386).
Ding, S. H., Fung, B., & Debbabi, M. (2015). A visualizable evidence-driven approach for authorship attribution. *ACM Transactions on Information and System Security, 17*(3), 1–30.
Ferrario, M. A., Simm, W., Whittle, J., Rayson, P., Terzi, M., & Binner, J. (2012). Understanding actionable knowledge in social media: Bbc question time and twitter, a case study.. In *Aaai conference on weblogs and social media* (pp. 455–458).
Fogués, R. L., Such, J. M., Espinosa, A., & Garcia-Fornes, A. (2014). Bff: A tool for eliciting tie strength and user communities in social networking services. *Information Systems Frontiers, 16*(2), 225–237.
Fong, P. W. (2011). Preventing sybil attacks by privilege attenuation: A design principle for social network systems. In *Ieee symposium on security and privacy* (pp. 263–278).
Garside, R. (1987). The claws word-tagging system. *The Computational Analysis of English: A Corpus-based Approach. London: Longman*, 30–41.
Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99*(12), 7821–7826.
Gupta, A., & Kumaraguru, P. (2012). Credibility ranking of tweets during high impact events. In *Proceedings of the workshop on privacy and security in online social media* (p. 2).
Halliday, J. (2011). London riots: How blackberry messenger played a key role. *The Guardian*.
Han, B., & Baldwin, T. (2011). Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the annual meeting of the association for computational linguistics: Human language technologies* (pp. 368–378).
Hankin, C. (2013). *Future identities: changing identities in the UK: the next 10 years. Final project report*. UK Parliament.
Hughes, D., Rayson, P., Walkerdine, J., Lee, K., Greenwood, P., Rashid, A., … Brennan, M. (2008). Supporting law enforcement in digital communities through natural language analysis. In *Computational forensics* (pp. 122–134).
Jackson, J. W., & Smith, E. R. (1999). Conceptualizing social identity: A new framework and evidence for the impact of different dimensions. *Personality and Social Psychology Bulletin, 25*(1), 120–135.
Kent, J. T. (1983). Information gain and a general measure of correlation. *Biometrika, 70*(1), 163–173.
Levine, M., & Crowther, S. (2008). The responsive bystander: how social group membership and group size can encourage as well as inhibit bystander intervention. *Journal of Personality and Social Psychology, 95*(6), 1429–1439.
Levine, M., & Koschate, M. (2014). Mums and feminists: Tracking shifts in identity salience in online environments. In *General meeting of the european association of social psychology* (p. 219).
Lustick, I. S. (2000). Agent-based modelling of collective identity: testing constructivist theory. *Journal of Artificial Societies and Social Simulation, 3*(1), 1.
Madejski, M., Johnson, M. L., & Bellovin, S. M. (2011). The failure of online social network privacy settings. *Technical Report CUCS-010-11*. Columbia University.
Matsuo, Y., Mori, J., Hamasaki, M., Nishimura, T., Takeda, H., Hasida, K., & Ishizuka, M. (2007). Polyphonet: an advanced social network extraction system from the web. *Web Semantics: Science, Services and Agents on the World Wide Web, 5*(4), 262–278.
Metz, C. E. (1978). Basic principles of roc analysis. In *Seminars in nuclear medicine: 8* (pp. 283–298). Elsevier.
Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C. R., & Song, D. (2012). On the feasibility of internet-scale author identification. In *Proceedings of the ieee symposium on security and privacy* (pp. 300–314).
Platow, M. J., Durante, M., Williams, N., Garrett, M., Walshe, J., Cincotta, S., … Barutchu, A. (1999). The contribution of sport fan social identity to the production of prosocial behavior.. *Group Dynamics: Theory, Research, and Practice, 3*(2), 161.
Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 248–256).
Rashid, A., Baron, A., Rayson, P., May-Chahal, C., Greenwood, P., & Walkerdine, J. (2013). Who am I? Analysing digital personas in cybercrime investigations. *IEEE Computer, 46*(4), 54–61.
Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., & Menczer, F. (2011). Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the international conference companion on world wide web* (pp. 249–252).
Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics, 13*(4), 519–549.
Reicher, S. D. (1996). The battle of Westminster: Developing the social identity model of crowd behaviour in order to explain the initiation and development of collective conflict. *European Journal of Social Psychology, 26*(1), 115–134.
Rousseau, D., & Van Der Veen, A. M. (2005). The emergence of a shared identity an agent-based computer simulation of idea diffusion. *Journal of Conflict Resolution, 49*(5), 686–712.
Smaldino, P., Pickett, C., Sherman, J., & Schank, J. (2012). An agent-based model of social identity dynamics. *Journal of Artificial Societies and Social Simulation, 15*(4), 7.
Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology, 60*(3), 538–556.
Stryker, S. (1980). *Symbolic interactionism: A social structural version*. Benjamin/Cummings Publishing Company.
Stryker, S., & Burke, P. J. (2000). The past, present, and future of an identity theory. *Social psychology quarterly*, 284–297.
Tajfel, H. (2010). *Social identity and intergroup relations*: 7. Cambridge University Press.
Taylor, V., Whittier, N., & Morris, A. (1992). Collective identity in social movement communities: Lesbian feminist mobilization. *Social perspectives in lesbian and gay studies (New York: Routledge, 1998)*, 349–365.
Thomas, K., McCoy, D., Grier, C., Kolcz, A., & Paxson, V. (2013). Trafficking fraudulent accounts: The role of the underground market in twitter spam and abuse.. In *Usenix security* (pp. 195–210).
Tufekci, Z., & Wilson, C. (2012). Social media and the decision to participate in political protest: Observations from tahrir square. *Journal of Communication, 62*(2), 363–379.
Verma, S., Vieweg, S., Corvey, W. J., Palen, L., Martin, J. H., Palmer, M., … Anderson, K. M. (2011). Natural language processing to the rescue?: Extracting'situational awareness' tweets during mass emergency. *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 385–392.
Wilson, A., & Rayson, P. (1993). Automatic content analysis of spoken discourse: a report on work in progress. *Corpus based computational linguistics*, 215–226.
Zhao, S., Grasmuck, S., & Martin, J. (2008). Identity construction on facebook: Digital empowerment in anchored relationships. *Computers in Human Behavior, 24*(5), 1816–1836.